Knowledge representation

lecture 3: processing natural language

T. Tammet, TUT

Kinds of NLP processing goals

There exist highly different methods for NLP, depending on the goals we have:

- Text similarity detection
- Automatic summarization
- Machine translation
- **Sentiment analysis**
- Speech processing
- **Information extraction**
- **Question answering**
- ...

NLP and restricted English

- By **NLP** people normally mean processing „generic“ unrestricted English: we will only partially understand the text and we will often misunderstand important parts.
- By **„restricted English“** people normally mean creating and processing a formal and restricted, yet „natural-looking“ English, with an exact meaning given to each sentence. Easy to understand, but really hard to write properly

Statistics and reasoning in NLP

- **Statistics and learning for creating probabilistic rules**

  It is very hard to capture and write down „language rules“:

  - There are far too many
  - Huge number of exceptions, exceptions to exceptions, ...
  - Everybody has her own version of language, and it changes all the time

- **Logic-based reasoning for disambiguation and knowledge extraction:**

  - Disambiguate the meaning of „she“, „this“, multi-meaning phrases etc
  - Convert parsed and annotated text to processable database form
  - Add general contextual knowledge
  - Answer questions about the text

Sentiment analysis

Why do sentiment analysis?

- Is this product review positive or negative?
- Is this customer email satisfied or dissatisfied?
- Based on a sample of tweets, how are people responding to this ad campaign/product release/news item?
- How have bloggers' attitudes about the president changed since the election?

For a detailed tutorial see

https://lct-master.org/files/MullenSentimentCourseSlides.pdf

Sentiment analysis

Example letter from a customer:

“Dear <hardware store>

Yesterday I had occasion to visit <your competitor>. The had an excellent selection, friendly and helpful salespeople, and the lowest prices in town.

You guys suck.

Sincerely,”

Sentiment analysis

Example amazon.com review: 1 star

„The original Star Wars trilogy was a defining part of my childhood. Born as I was in 1971, was just the right age to fall headlong into this amazing new world Lucas created. I was one of those kids that showed up early at toy stores [...] anxiously awaiting each subsequent installment of the series. I'm so glad that by my late 20s, the old thrill had faded, or else I would have been EXTREMELY upset over Episode I: The Phantom Menace ... perhaps the biggest let-down in film history.”

Sentiment analysis

Brief notes on techniques:

- Get a collection of **positive/negative phrases** with pos/neg scores like „excellent“, „fine“, „nasty“ etc
- Get a collection of **score incrementing / decrementing** words like „very“, „a bit“, ...
- Do shallow grammar analysis for **inversion** like „not ....“
- Train your phrase database on already tagged/weighted text co llections like amazon or imdb reviews

Sentiment analysis

Try out this demo, built using Python NLTK:

http://text-processing.com/demo/sentiment/

Grammar: detect sentence structure

Different ways to go:

- „Linguistic“ grammar based on word types
- „Semantic“ grammar based on word meanings

Grammar: detect sentence structure

Typical syntactic categories used in NLP:

s – sentence „fruit flies like a rotten apple“

np - noun phrase „fruit flies“

vp - verb phrase „like a rotten apple“

det - determiner (article) „a“

n – noun „fruit“

tv - transitive verb (takes an object) „like“

iv - intransitive verb (a la „birds **fly**“, „I **sneezed**“)

prep – preposition „at“, „with“, „in“, „to“, „on“, ...

pp - prepositional phrase (a la „at home“, „with me“)

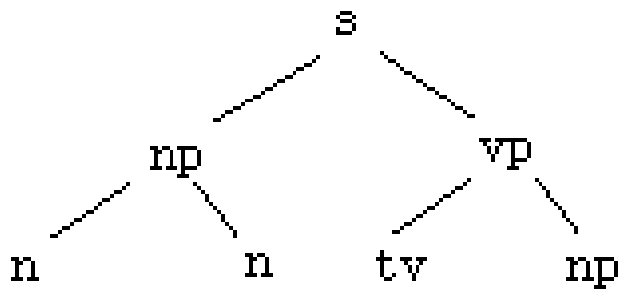
adj - adjective „rotten“

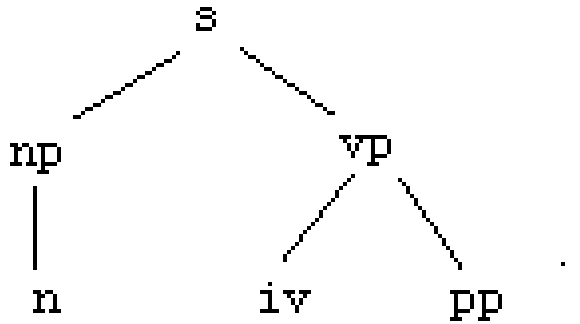
Grammar: detect sentence structure

Many alternative grammatical readings: which to choose?

Figure 2. An ambiguous grammar and partial parse trees for "fruit flies like an apple."

```
s    ⟶ np vp
np   ⟶ det n
     ⟶ n
     ⟶ n n
vp   ⟶ tv np
     ⟶ iv pp
pp   ⟶ prep np
det  ⟶ a
     ⟶ an
n    ⟶ fruit
     ⟶ apple
     ⟶ flies
iv   ⟶ flies
tv   ⟶ like
prep ⟶ like
```





Grammar: detect sentence structure

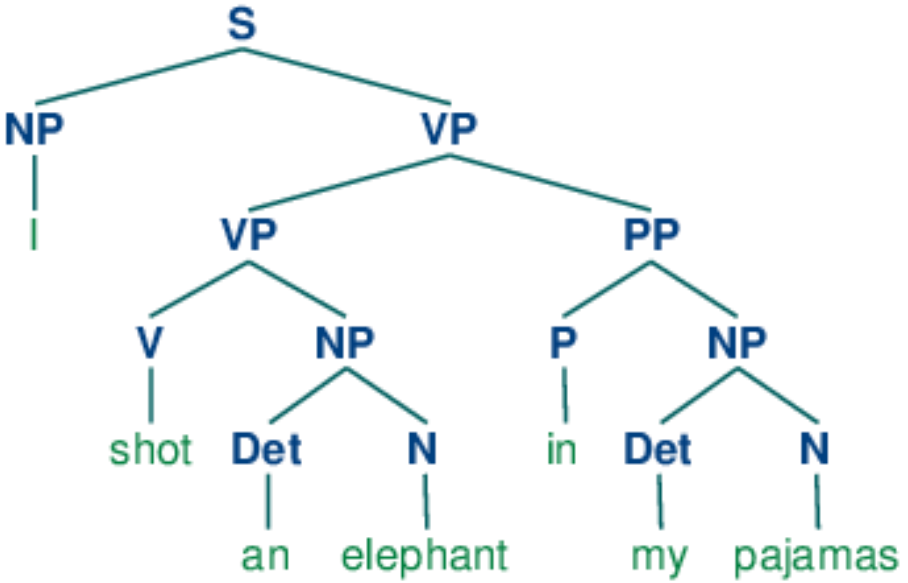
Part-of-speech (POS) tagging:

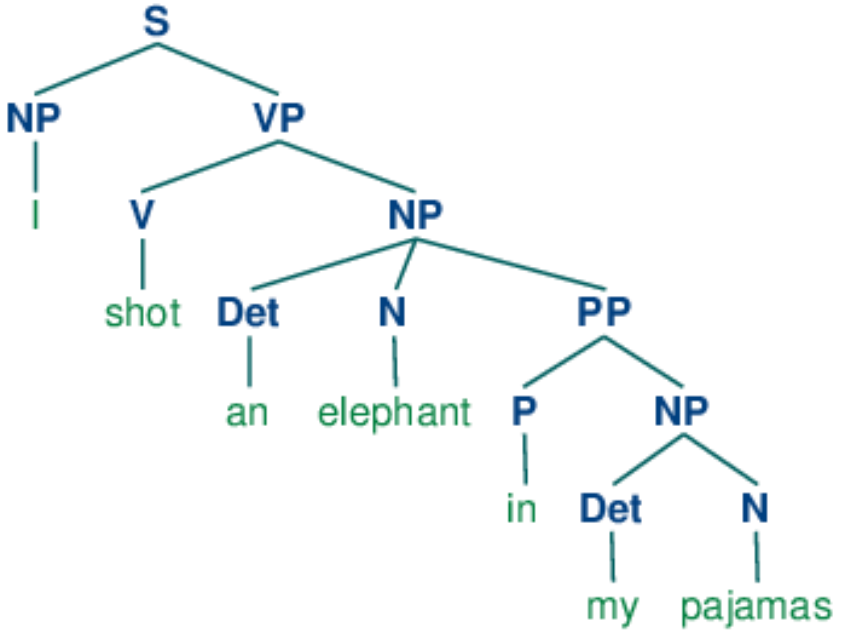
Try out http://smile-pos.appspot.com/

Fruit/NNP flies/VBZ like/IN a/DT rotten/JJ apple/NN ./.

Ambiguity in grammar

I shot an elephant in my pajamas.





Semantic parsing

Citation:

> „Semantic parsing is the process of mapping a natural-language sentence into a formal representation of its meaning.
>
> A shallow form of semantic representation is a case-role analysis (a.k.a. a semantic role labeling), which identifies roles such as agent, patient, source, and destination.
>
> A deeper semantic analysis provides a representation of the sentence in predicate logic or other formal language which supports automated reasoning.“

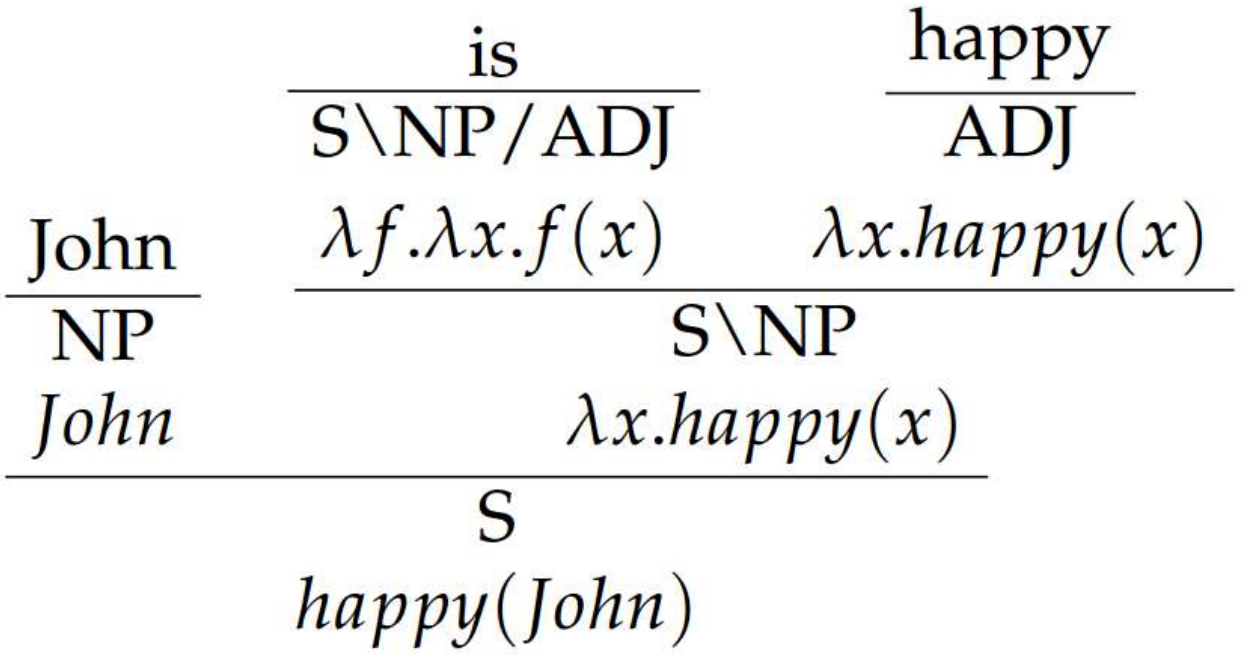
Have a look at the thorough tutorial

http://www.lsi.upc.edu/~ageno/anlp/semanticParsing.pdf

Semantic parsing

One approach: **words need parameters**, like functions.

John is happy:



Why Watson Won

Jim Hendler

(and Simon Ellis)

Tetherless World Professor of Computer,Web and Cognitive Sciences
Director, Rensselaer Institute for Data Exploration and Applications

Rensselaer Polytechnic Institute (RPI)

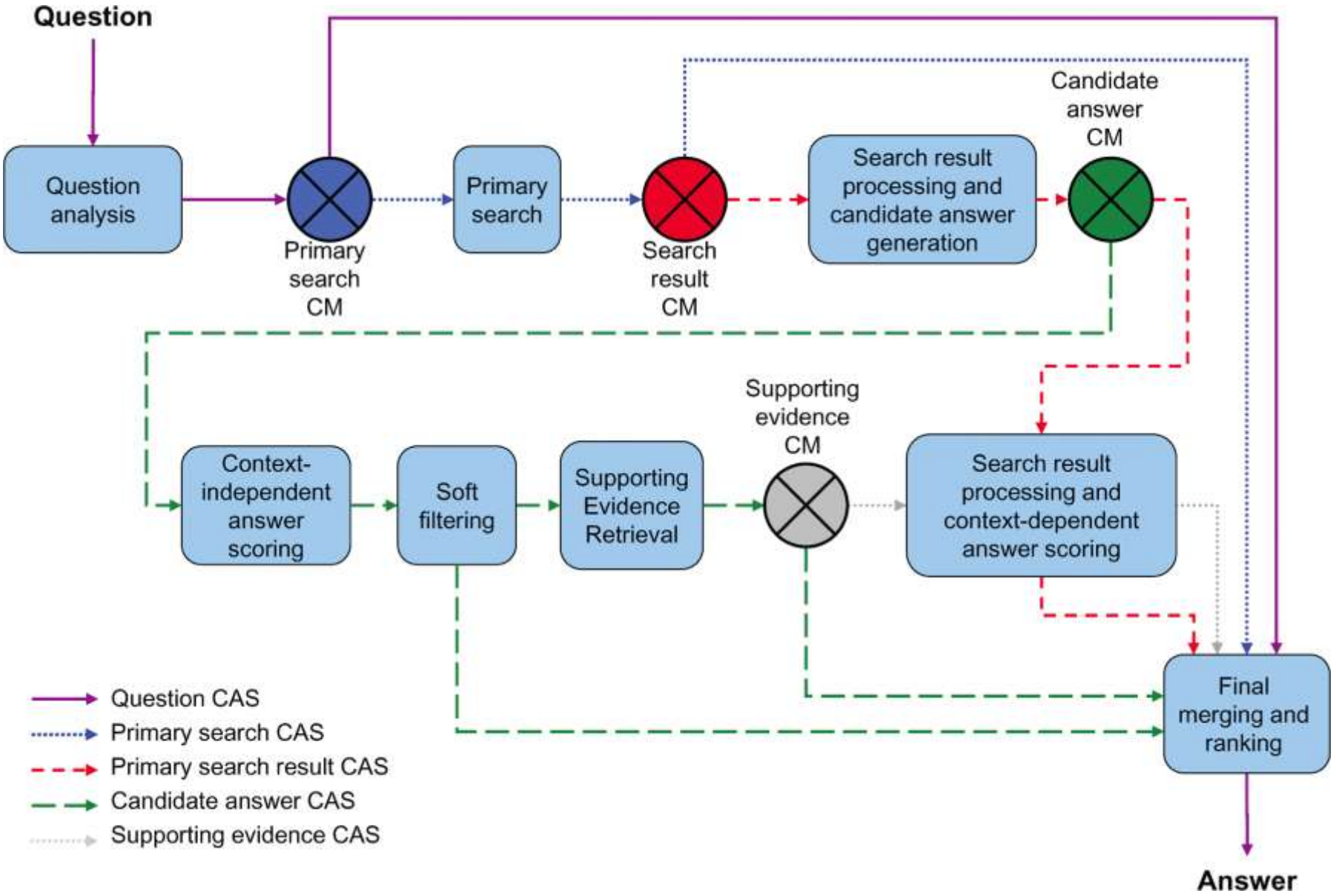
http://www.cs.rpi.edu/~hendler

@jahendler (twitter)

IBM Watson

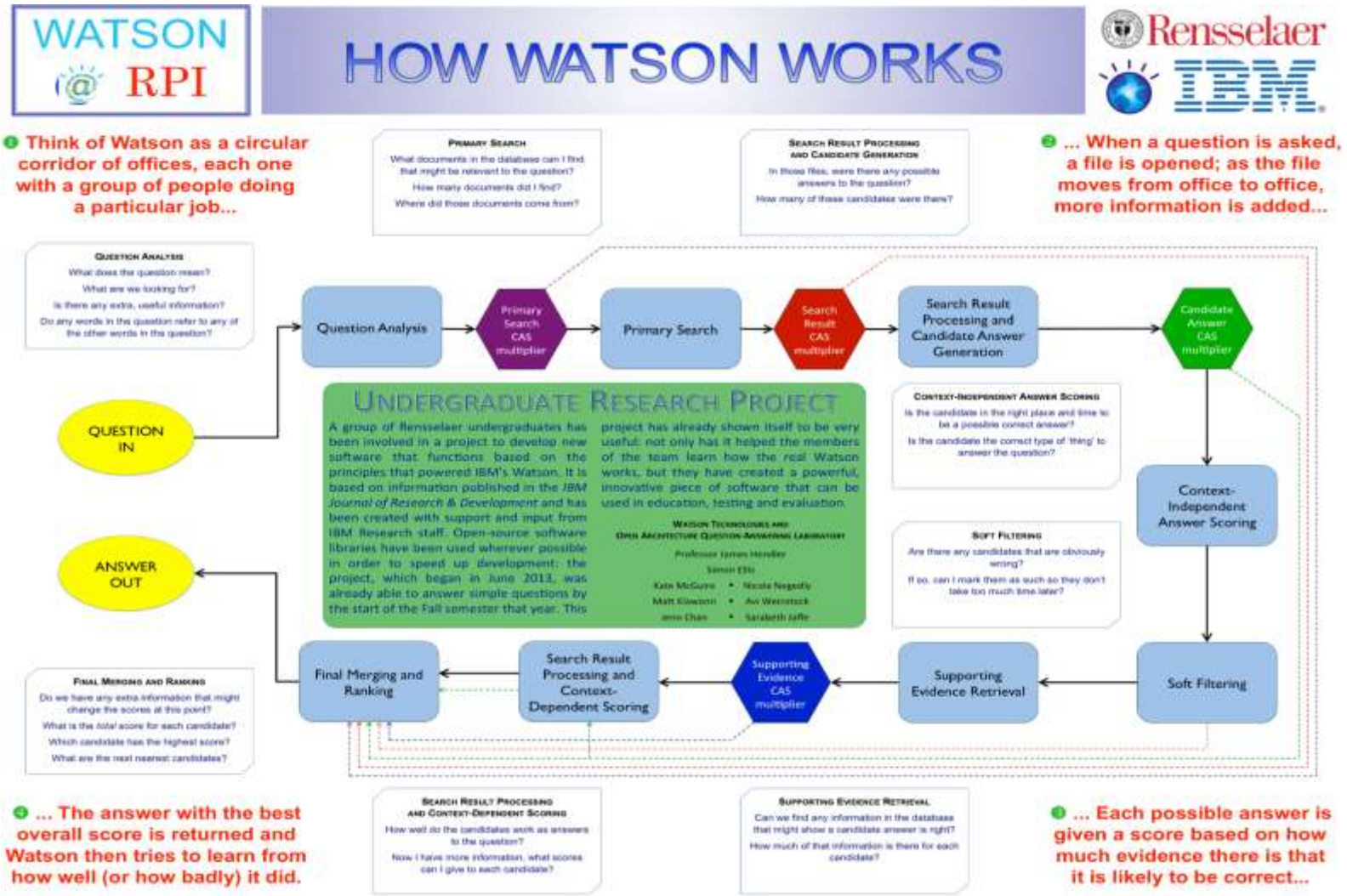

THINK
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Inside Watson

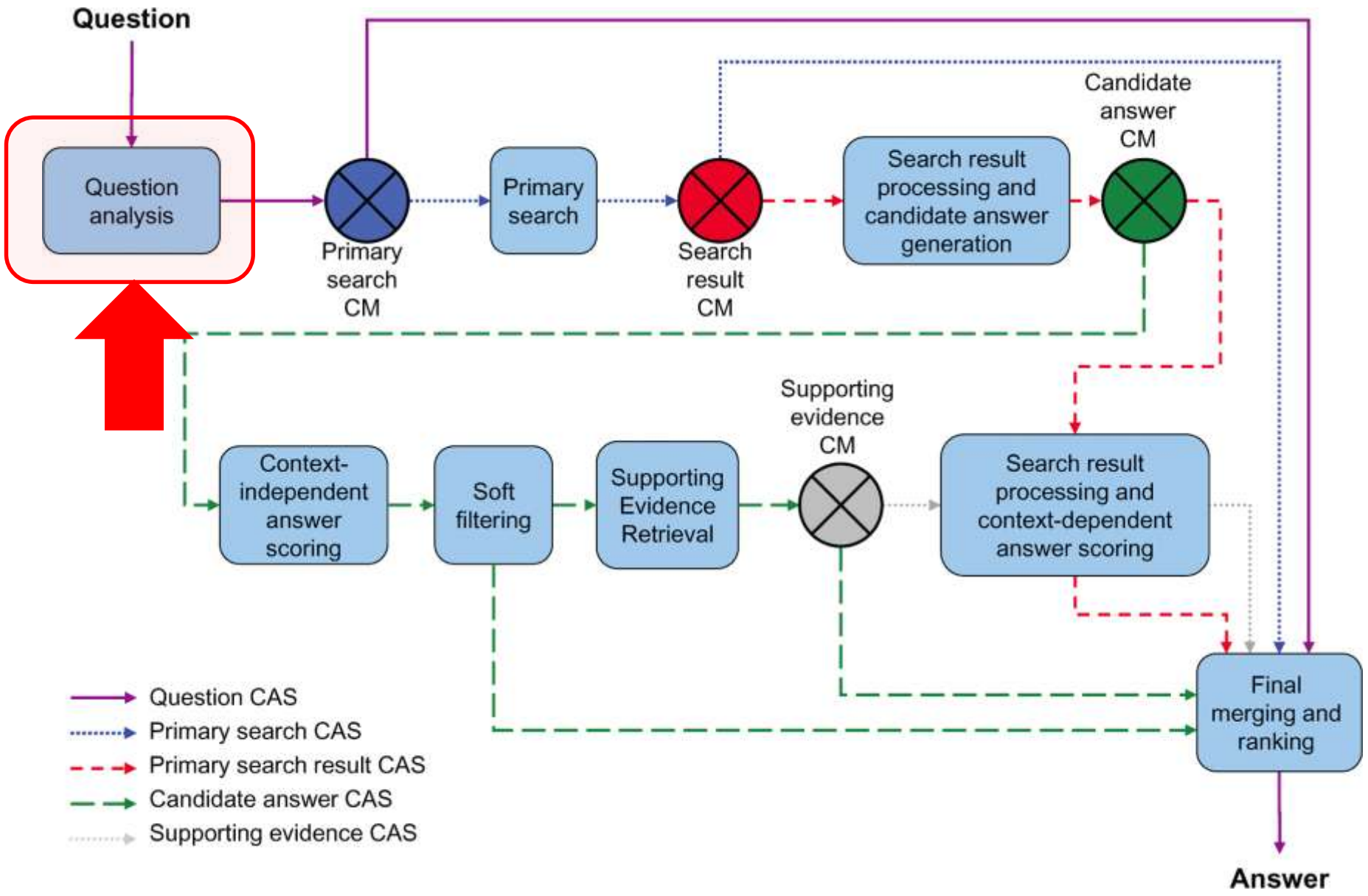


Watson pipeline as published by IBM; see *IBM J Res & Dev* **56** (3/4), May/July 2012, p. 15:2

Watson Simplified (S. Ellis, 2013)



Question Analysis

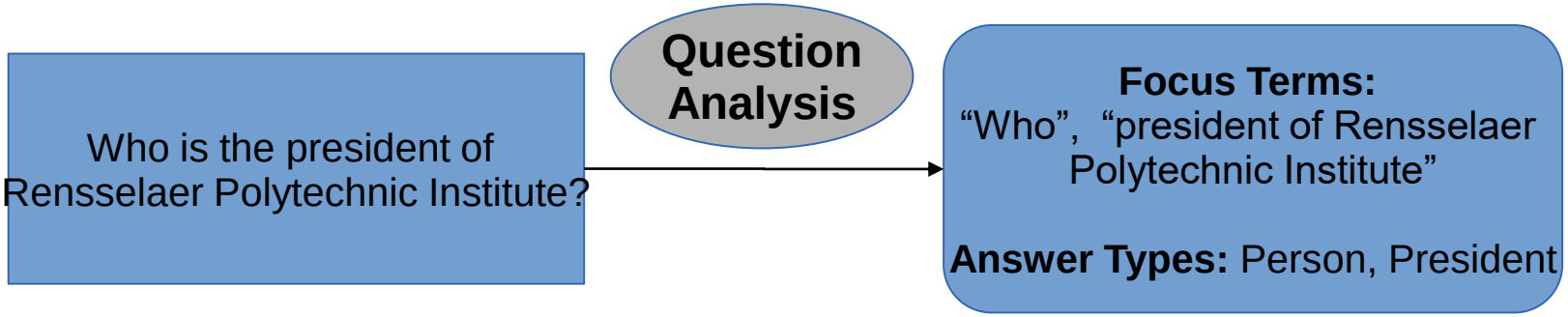


Question analysis

What is the question asking for?

Which terms in the question refer to the answer?

Given any natural language question, how can Watson accurately discover this information?



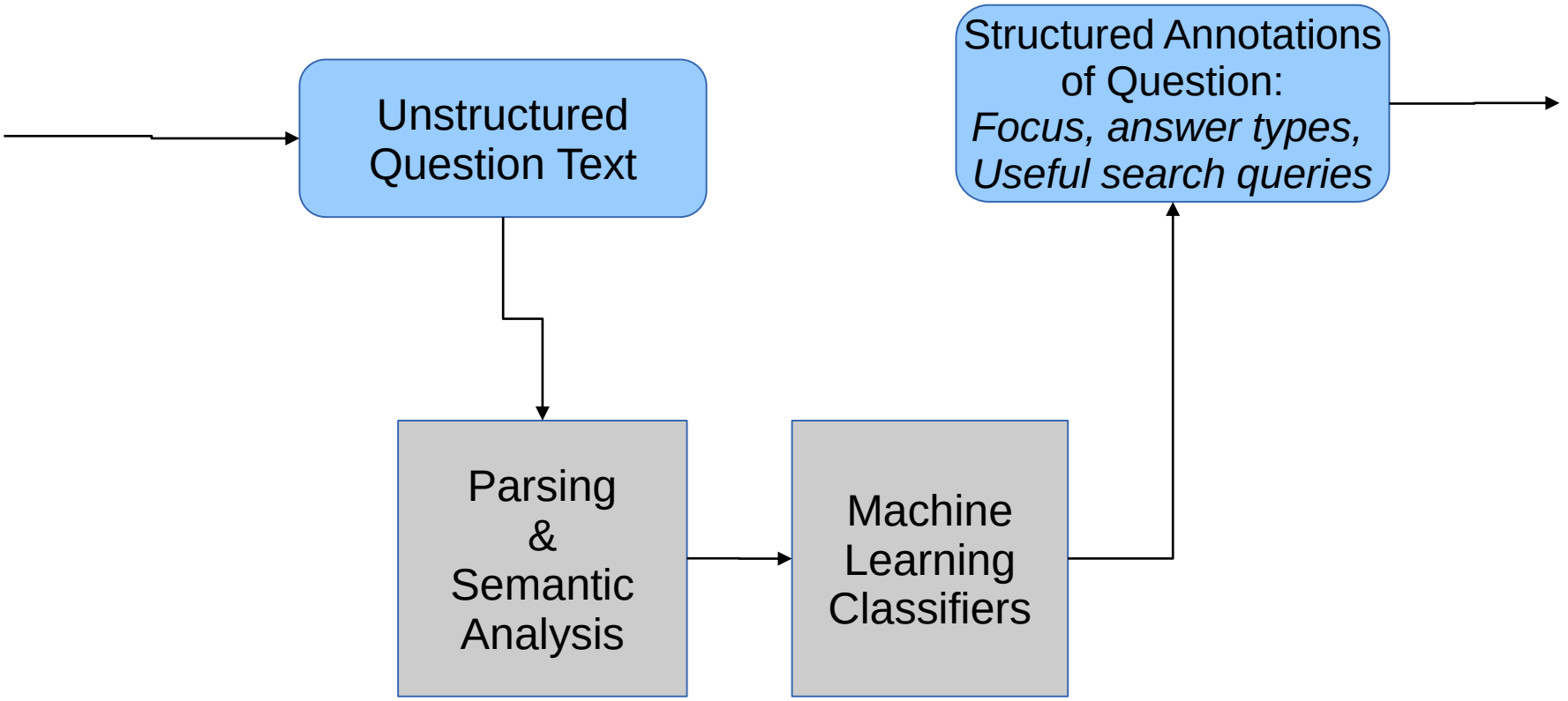
Parsing and semantic analysis

What information about a previously unseen piece of English text can Watson determine?

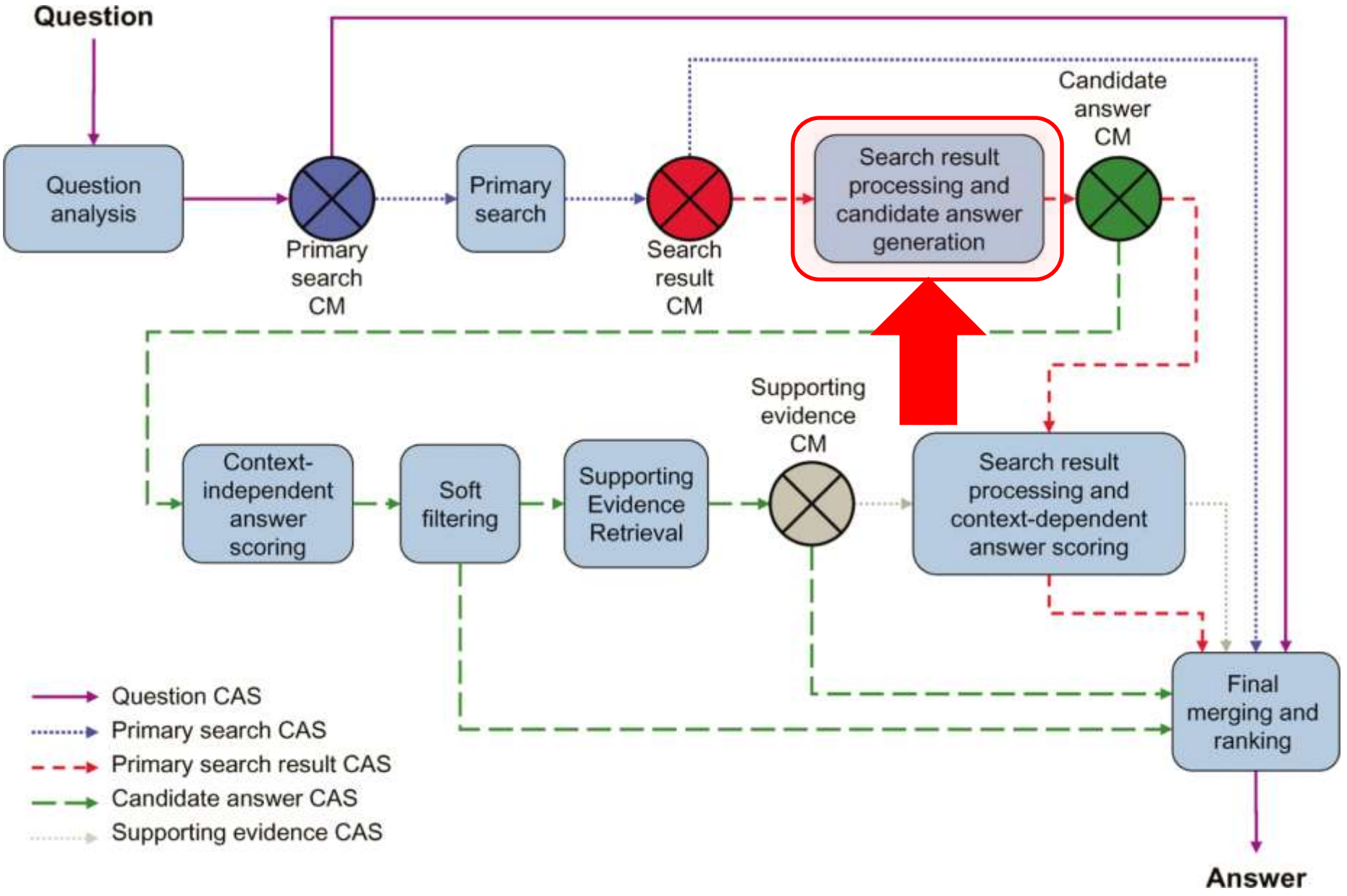
How is this information useful?

Natural Language Parsing	Semantic Analysis
- *grammatical* structure - parts of speech - relationships between words - ...*etc.*	- *meanings* of words, phrases, etc. - synonyms, entailment - hypernyms, hyponyms - ...*etc.*

Question analysis pipeline



Search Result Processing and Candidate Generation



Primary Search

Primary Search is used to generate the corpus of information from which to take candidate answers, passages, supporting evidence, and essentially all textual input to the system

It formulates queries based on the results of Question Analysis

These queries are passed into a (cached) search engine which returns a set number of highly relevant documents and their ranks.

Candidate Generation

Candidate Generation generates a wide net of possible answers for the question from each document.

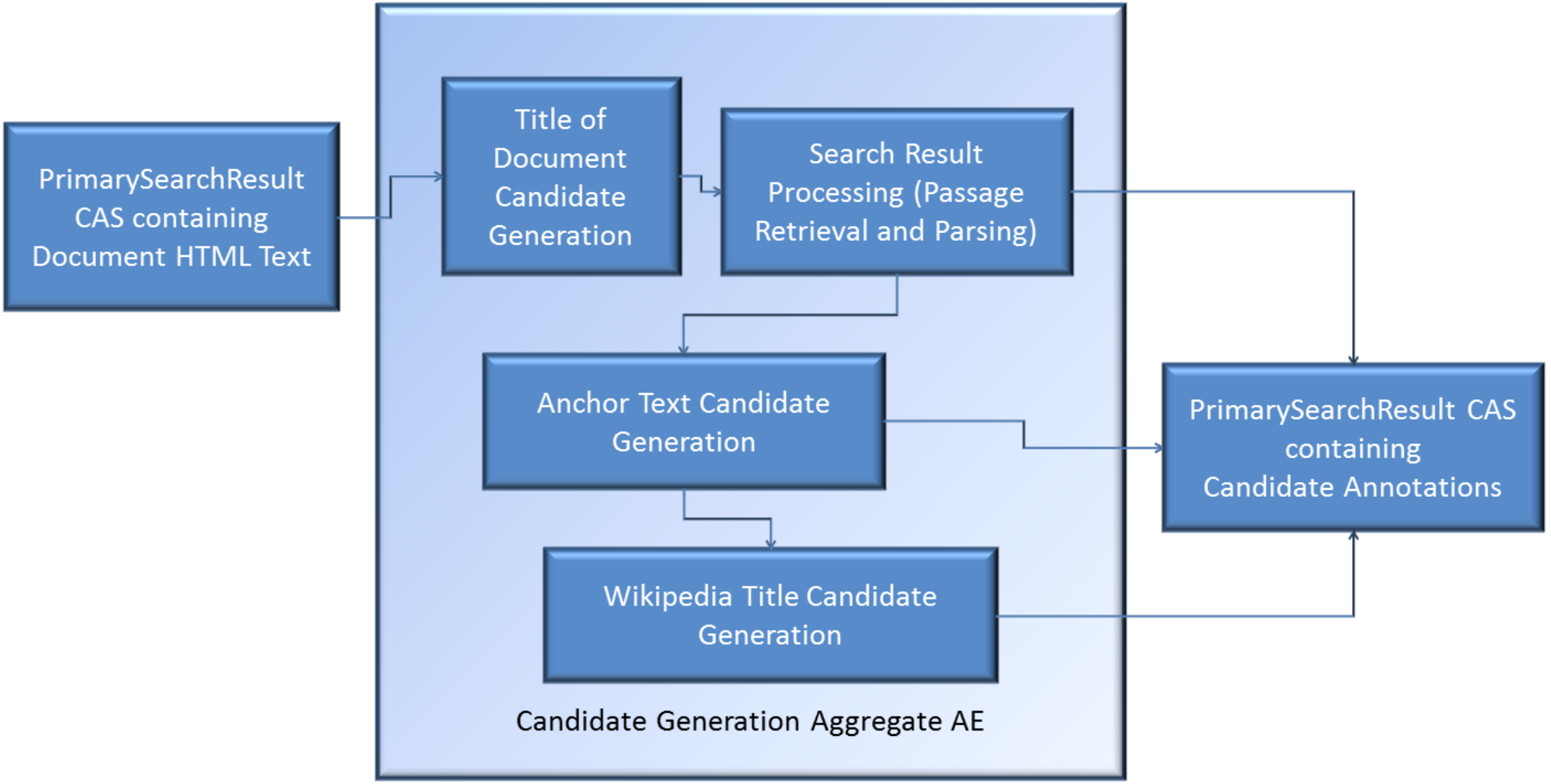
Using each document, and the passages created by Search Result Processing, we generate candidates using three techniques:

Title of Document (T.O.D.): Adds the title of the document as a candidate.

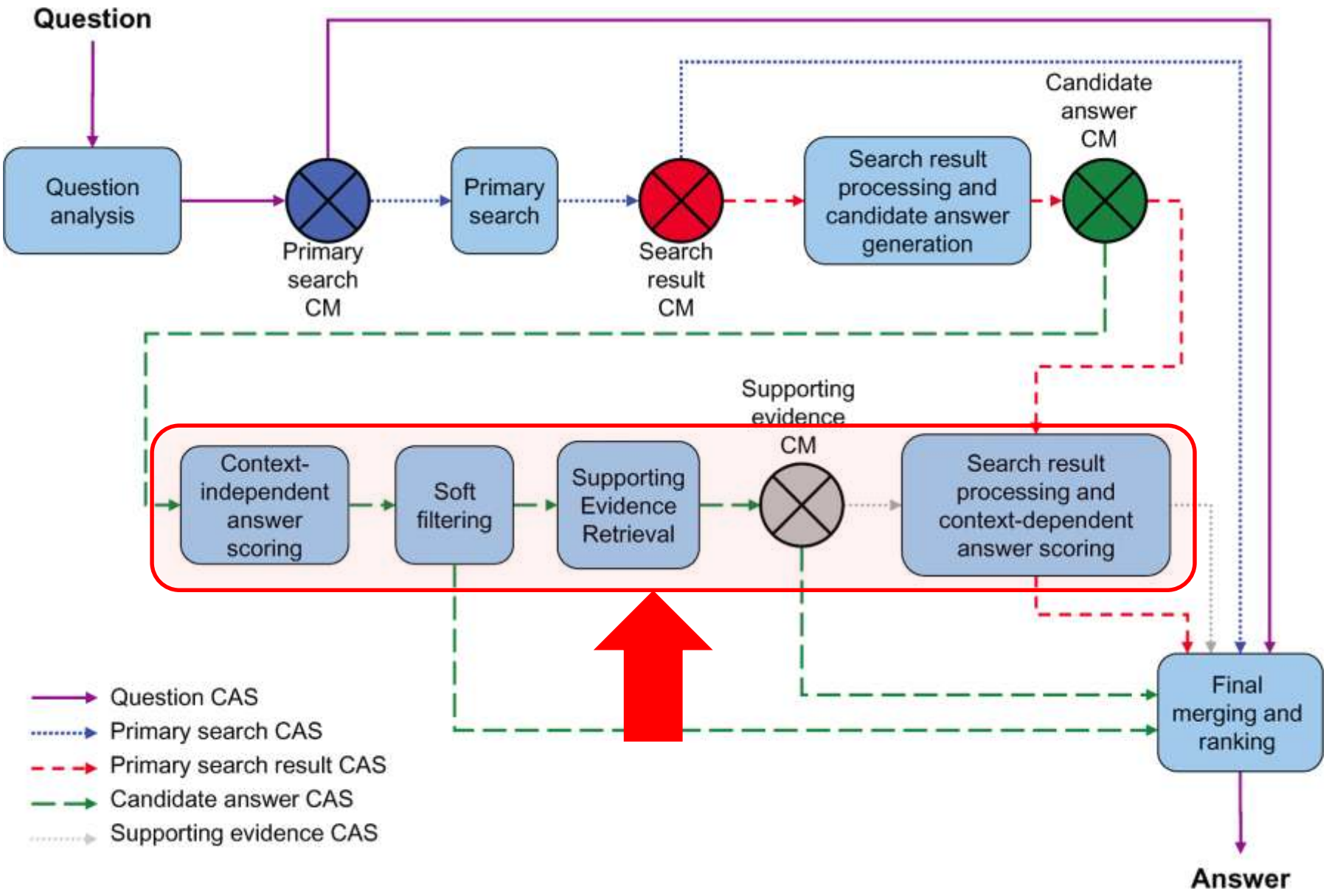
Wikipedia Title Candidate Generation: Adds any noun phrases within the document's passage texts that are also the titles of Wikipedia articles.

Anchor Text Candidate Generation: Adds candidates

Search Result Processing and Candidate Generation



Scoring & Ranking



Scoring

Analyzes how well a candidate answer relates to the question

Two basic types of scoring algorithm

- Context-independent scoring
- Context-dependent scoring

Types of scorers

- Context-independent
  - Question Analysis
  - Ontologies (DBpedia, YAGO, etc)
  - Type hierarchy reasoning
- Context-dependent
  - Analyzes feature of the natural language environment where candidates were found
    - Relies on “passages” found during search
  - Many special purpose ones used in Jeopardy

Scorers

Passage Term Match

Textual Alignment

Skip-Bigram

- Each of these scores supportive evidence
- These scores are then merged to produce a single candidate score

Example:Textual Alignment

Finds an optimal alignment of a question and a passage

Assigns “partial credit” for close matches

“Who is the President of RPI?”

Who President of RPI.

Shirley Ann Jackson is the President of RPI.

Skip-Bigram

Constructs a graph

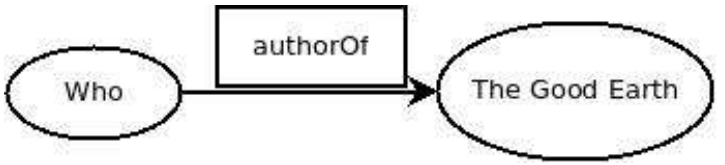
- Nodes represent terms (syntactic objects)
- Edges represent relations

Extracts skip-bigrams

- A skip-bigram is a pair of nodes either directly connected or which have only one intermediate node
- Skip-bigrams represent close relationships between terms

Example

Who authored
"The Good Earth"?



"Pearl Buck, author of
the good earth…"

Watson Summary

- Watson works by
  - Analyzing the question
    - natural language parsing
    - text extraction
  - Generating a large number of candidates
    - mostly search heuristics
  - Scoring each
    - through multiple scorers
    - with weights adjusted by learning algorithm
  - Returning top candidate

Useful additional links

Some notes here:

https://courses.cs.ttu.ee/pages/Useful_NLP_links_and_notes

Starting with Jurafsky/Manning Stanford course:

- NER

  http://spark-public.s3.amazonaws.com/nlp/slides/Information_Extraction_and_Named_Entity_Recognition_v2.pptx

- POS tagging

  http://spark-public.s3.amazonaws.com/nlp/slides/Maxent_PosTagging.pptx

- Parsing

  http://spark-public.s3.amazonaws.com/nlp/slides/Parsing-Intro.pptx

Then

http://www.cogsci.rpi.edu/courses/**logic**andai/documents/LCL.ppt